A review of asymptotic theory of estimating functions

Jacod, Jean ; Sørensen, Michael

Published in:
Statistical Inference for Stochastic Processes

DOI:
10.1007/s11203-018-9178-8

Publication date:
2018

Document version
Early version, also known as pre-print

A review of asymptotic theory of estimating functions

Jean Jacod
Institut de Mathématiques de Jussieu and
Université P. et M. Curie (Paris-6)
CNRS UMR 7586
75252 Paris Cédex 05
France

Michael Sørensen
Dept. of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark

Abstract

Asymptotic statistical theory for estimating functions is reviewed in a generality suitable for stochastic processes. Conditions concerning existence of a consistent estimator, uniqueness, rate of convergence, and the asymptotic distribution are treated separately. Our conditions are not minimal, but can be verified for many interesting stochastic process models. Several examples illustrate the wide applicability of the theory and why the generality is needed.

Key words: Asymptotic statistics, diffusion processes, ergodic processes, high frequency asymptotics, limit theory, longitudinal data, Markov process, misspecified model.

1 Introduction

Estimating functions provide a general framework for finding estimators and studying their properties in many different kinds of statistical models and asymptotic scenarios. We present the main results of the asymptotic theory of estimating functions in a generality that is suitable for statistical inference for stochastic processes. Typically, we have observations of n random variables, $X_1, \ldots, X_n$, whose joint law depends on a parameter θ. Also, for instance, the sampling frequency or the diffusion coefficient of a diffusion process may depend on n. An estimating function is a function of the data as well as the parameter, $G_n(\theta) = G_n(\theta, X_1, \ldots, X_n)$, and estimators are obtained by solving the estimating equation $G_n(\theta) = 0$. We present simple conditions for existence, uniqueness, consistency, rate of convergence and asymptotic distribution of such estimator as n tends to infinity. Our conditions are not minimal, but can be verified for many interesting stochastic process models. This is illustrated by several examples, which also demonstrate why the generality of our conditions is needed.

The theory covers consistent estimators obtained by maximising (or minimising), with respect to θ, a function $H_n(\theta, X_1, \ldots, X_n)$ of the data and the parameter. If H_n is continuously differentiable, and if the true parameter value belongs to the interior of the parameter space, then eventually the estimator will solve the estimating equation $\partial_\theta H_n(\theta) = 0$. This includes, of course, maximum likelihood estimation. It is often the case for stochastic process models that the likelihood function is not explicitly available, is not tractable, or requires heavy computations. In such cases other estimating functions provides a tractable alternative.

The idea of using estimating equations is an old one and goes back at least to Karl Pearson's introduction of the method of moments. The term estimating function may have been coined by Kimball (1946). In the statistics literature, the modern theory of optimal estimating functions dates back to the papers by Godambe (1960) and Durbin (1960), however the basic idea was in a sense already used in Fisher (1935). The theory was extended to stochastic processes by Godambe (1985), Godambe and Heyde (1987), and several others; see the references in Heyde (1997). In the econometrics literature the foundation was laid by Hansen (1982). This paper treats a general class of models including time series models. Asymptotic theory was developed in, e.g., Hansen (1982, 1985), Chamberlain (1987), and Kuersteiner (2002). The theory was extended by several authors, see the discussion and the discussion and references in Hall (2005). A discussion of links between the econometrics and statistics literature can be found in Hansen (2001).

The classical approach to asymptotic statistical theory for estimating functions is based on the seminal work of Cramér (1946). To prove asymptotic existence of an estimator, one approach, originally due to Aitchison and Silvey (1958), is based on Brouwer's fixed-point theorem. This idea is used in, e.g., Sweeting (1980), Barndorff-Nielsen and Sørensen (1994), and Sørensen (1999). In the present paper, we follow instead an approach based on the fixed-point theorem for contractions, which was also used by, e.g., Yuan and Jennrich (1998). The latter approach gives relatively simple conditions and a straightforward identifyability condition that implies a uniqueness result for estimators. A comprehensive introduction to asymptotic statistics can be found in van der Vaart (1998).

The paper is organised as follows. In Section 2 we present the general asymptotic theory, while theory for ergodic processes is presented in Section 3. Longitudinal data are briefly considered in Section 4, and high frequency data, both with a fixed and an infinite time horizon, are discussed in Section 5. All proofs are collected in Section 6.

2 Asymptotics for estimating functions

The general set-up is as follows. We have a measurable space $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}$, the *true measure*, and a family $(\mathbb{P}^{(\theta)})_{\theta\in\Theta}$ of probability measures indexed by $\Theta \subseteq \mathbb{R}^p$, the *statistical model*. Often it is assumed that there exists a $\theta_0 \in \Theta$ (called the true parameter value) such that $\mathbb{P} = \mathbb{P}^{(\theta_0)}$, but this need not be the case.

At stage n, we have a set of observations which generates a $\mathbb{P}$–complete σ–field $\mathcal{F}_n$ (that is, $\mathcal{F}_n$ is the $\mathbb{P}$–completion of the σ–field $\mathcal{F}_n^0$ generated by those observations.) Taking the completion simplifies the mathematical formulation of the results below, and it has no negative consequences from a practical viewpoint since all statements about estimators are always up to a $\mathbb{P}$-null set. To be consistent, we also suppose that $\mathcal{F}$ is $\mathbb{P}$–complete. If we observe a continuous time stochastic process X_t, then $\mathcal{F}_n$ could, for instance, be the (complete) σ-field generated by the variables X_s for $s \in [0, t_n]$ for some increasing sequence t_n, or by the variables $X_{i\Delta_n}$ for $i = 0, \ldots, n$ and some $\Delta_n > 0$.

An *estimating function* at stage n is a function $(\theta, \omega) \mapsto G_n(\theta, \omega)$ that takes its values in $\mathbb{R}^p$ and depends on the statistical parameter $\theta \in \Theta$ and on the observation at this stage, that is $(\theta, \omega) \mapsto G_n(\theta, \omega)$ is measurable w.r.t. the product of the Borel σ-field of Θ with $\mathcal{F}_n$. For convenience, we usually suppress the argument ω in the notation. We get an estimator $\widehat{\theta}_n$ by solving the estimating equation

$$G_n(\theta) = 0. \tag{2.1}$$

For any given ω, this equation may have a unique solution, several solutions, or none, so we have to be a bit careful. Therefore we give the following formal definition, where δ denotes a "special" point, which we take to be outside Θ because this is most convenient, and $\Theta_\delta = \Theta \cup \{\delta\}$.

Definition 2.1 *a) The* domain of definition *of G_n-estimators (for a given n) is the set D_n of all ω for which $G_n(\theta, \omega) = 0$ for at least one $\theta \in \Theta$.*

b) A G_n-estimator*, $\widehat{\theta}_n$, is any $\mathcal{F}_n$–measurable map from Ω into Θ_δ, such that $\widehat{\theta}_n \in \Theta$ and $G_n(\widehat{\theta}_n) = 0$ on the set D_n and $\widehat{\theta}_n = \delta$ on its complement $(D_n)^c$.*

Because $\mathcal{F}_n$ is $\mathbb{P}$-complete, the measurable selection theorem implies that $A_n \in \mathcal{F}_n$ and that a G_n–estimator always exists.

In the rest of this paper, $\mathcal{M}_p$ denotes the set of all $p \times p$ matrices, and $\mathcal{M}_p^{\text{inv}}$ the subset of all $A \in \mathcal{M}_p$ that are invertible (non–singular). A G_n–estimator is also a B_nG_n–estimator for any $B_n \in \mathcal{M}_p^{\text{inv}}$, where B_n can depend on the data as well as the parameter. We call the family of estimating functions of the form B_nG_n *versions* of G_n. Since B_n depends on n, the assumptions about the asymptotic behavior of G_n made in the following can obviously fail for some versions of G_n, but it is sufficient that a version exists, for which the conditions are satisfied. All limits below are taken as $n \to \infty$.

2.1 Existence and uniqueness of a consistent estimator

The following condition ensures that for n large enough the estimating equation has a solution that converges to a particular parameter value $\bar{\theta}$. When the statistical model contains the true model, the estimating function should preferably be chosen such that $\bar{\theta} = \theta_0$. It is, however, useful to include the more general case in the theory. A couple of exemples below will illustrate why.

Condition 2.2 *There exists a parameter value $\bar{\theta} \in \operatorname{int}\Theta$ (the interior of Θ), a neighbourhood M of $\bar{\theta}$, and a (possibly random) $\mathbb{R}^p$–valued function G on M, such that the following holds:*

(i) $G_n(\bar{\theta}) \xrightarrow{\mathbb{P}} 0$ *(convergence in probability under the true measure $\mathbb{P}$) and $G(\bar{\theta}) = 0$.*

(ii) *For $\mathbb{P}$-almost all ω, $G(\cdot, \omega)$ and all $G_n(\cdot, \omega)$ are C^1 (i.e., continuously differentiable) on M, and*

$$\sup_{\theta \in M} \| \partial_\theta G_n(\theta) - \partial_\theta G(\theta) \| \xrightarrow{\mathbb{P}} 0. \tag{2.2}$$

(iii) *The matrix $\partial_\theta G(\bar{\theta})$ is non-singular with $\mathbb{P}$–probability one.*

In (2.2) $\| \cdot \|$ could be any norm on $\mathcal{M}_p$, since any two norms on a finite-dimensional space are equivalent. However, in the following we will use the operator norm

$$\|A\|^2 = \max\{|\lambda_j| : \lambda_j \text{ is an eigenvalue of } A^*A\},$$

or equivalently $\|A\| = \sup_{|x|=1} |Ax|$. Here and below $|\cdot|$ denotes the Euclidian norm, and A^* is the transpose of A. The vector $\bar{\theta}$ and the function $G(\theta)$ depend on the true probability measure $\mathbb{P}$. Note that when M is a bounded set, (2.2) implies

$$\sup_{\theta \in M} |G_n(\theta) - G(\theta)| \xrightarrow{\mathbb{P}} 0. \tag{2.3}$$

It is not a restriction to assume in Condition 2.2 that M is convex, or even an open or closed ball centered at $\bar{\theta}$, since M always contains such a ball. Since G_n and G are C^1 on M outside a $\mathbb{P}$-null set, whereas $\mathcal{F}$ is $\mathbb{P}$-complete, the left side of (2.2) is $\mathcal{F}$–measurable. Moreover, it is not necessary to assume the existence of the function G in the condition, where $\partial_\theta G(\theta)$ can be replaced by a function $W(\theta) \in \mathcal{M}_p$ such that $W(\bar{\theta}) \in \mathcal{M}_p^{\text{inv}}$. Then the existence of a G such that $\partial_\theta G(\theta) = W(\theta)$ and $G(\bar{\theta}) = 0$ follows.

At this point it is not obvious why we do not simply take $\bar{\theta}$ to be equal to the true parameter value θ_0. Before presenting the asymptotic theory, let us therefore present two simple examples showing that there are situations where quite naturally $\bar{\theta} \neq \theta_0$. Note that the theory below would in no way be simplified by assuming that $\bar{\theta} = \theta_0$.

Example 2.3 Suppose we model observations $X_0, X_1, \ldots, X_n$ by the autoregression of order one

$$X_i = \theta X_{i-1} + \epsilon_i, \tag{2.4}$$

where the ϵ_i's are i.i.d. random variables with mean zero and finite variance, and $\theta \in (-1, 1)$ so that X is ergodic. It is natural to estimate θ by minimizing

$$H_n(\theta) = n^{-1} \sum_{i=1}^{n} (X_i - \theta X_{i-1})^2,$$

which is minus the logarithm of a Gaussion pseudo-likelihood. This least squares estimator can be found by solving the estimating equation $G_n(\theta) = 0$, where

$$G_n(\theta) = n^{-1} \sum_{i=1}^{n} X_{i-1}(X_i - \theta X_{i-1}).$$

If our observations are in fact generated by (2.4) with $\theta = \theta_0$, then $G_n(\theta_0) = n^{-1} \sum_{i=1}^{n} X_{i-1}\epsilon_i \xrightarrow{\mathbb{P}} 0$ by the law of large numbers for martingales. It is not difficult to see that θ_0 is the only parameter value for which $G_n(\theta) \xrightarrow{\mathbb{P}} 0$, so necessarily $\bar{\theta} = \theta_0$.

Now assume that our data are actually observations from an autoregression of order two, i.e. that

$$X_i = \theta_1 X_{i-1} + \theta_2 X_{i-2} + \epsilon_i,$$

where the ϵ_i's are as before, and where θ_1 and θ_2 are such that the observed process X is ergodic. Thus if $\theta_2 \neq 0$, our statistical model is misspecified. When this is the case,

$$G_n(\theta) = (\theta_1 - \theta) n^{-1} \sum_{i=1}^{n} X_{i-1}^2 + \theta_2 n^{-1} \sum_{i=1}^{n} X_{i-1} X_{i-2} + n^{-1} \sum_{i=1}^{n} X_{i-1}\epsilon_i,$$

implying that

$$G_n(\theta) \xrightarrow{\mathbb{P}} (\theta_1 - \theta)\sigma^2 + \theta_2 \nu,$$

where σ^2 and ν are the expectations of X_i^2 and $X_{i-1}X_i$ under the stationary distribution for X. We see that necessarily $\bar{\theta} = \theta_1 + \theta_2 \nu / \sigma^2$. The least squares estimator converges to $\bar{\theta}$ as $n \to \infty$ when the observed process X is an autoregression of order two. □

Example 2.4 As another closely related example, consider the Ornstein-Uhlenbeck process given by

$$dX_t = -\theta X_t dt + dW_t, \quad \theta > 0. \tag{2.5}$$

Suppose we have observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ from (2.5) with $\theta = \theta_0$. The parameter θ is often estimated by the least squares estimator obtained by minimizing the function

$$H_n(\theta) = n^{-1} \sum_{i=1}^{n} (X_{i\Delta} - (1 - \theta\Delta) X_{(i-1)\Delta})^2$$

This is motivated by an Euler discretization and is the same as using the estimating function

$$G_n(\theta) = n^{-1} \sum_{i=1}^{n} X_{(i-1)\Delta} [X_{i\Delta} - (1 - \theta\Delta) X_{(i-1)\Delta}].$$

Since $X_{i\Delta} = e^{-\theta_0 \Delta} X_{(i-1)\Delta} + \epsilon_i$, where the ϵ_i's satisfy the assumptions in Example 2.3, it follows from this example that $(1 - \bar{\theta}\Delta) = e^{-\theta_0 \Delta}$ or $\bar{\theta} = (1 - e^{-\theta_0 \Delta})/\Delta$. The mean squares estimator $\hat{\theta}_n$ obtained by solving $G_n(\theta) = 0$ converges to $\bar{\theta}$ as $n \to \infty$. When $\theta_0 \Delta$ is small, $\bar{\theta}$ is close to θ_0, but if $\theta_0 \Delta$ is large, $\bar{\theta}$ will be very far from θ_0. □

Condition 2.2 implies that $\bar{\theta}$ is a.s. an isolated zero of G. In view of (2.3), it is intuitively clear that G_n must have a unique zero near $\bar{\theta}$ when n is sufficiently large. That is the content of the following theorem. When θ is one-dimensional, it is easy to turn the intuitive argument into a rigorous proof. When $p > 1$, this is most easily done using a fixed point theorem as we shall do in the proof of Theorem 2.5.

A sequence of estimators that converges to a parameter value different from the true value is usually not called consistent, but to facilitate the discussion, we call a sequence $(\widehat{\theta}_n)$ of estimators *weakly $\bar{\theta}$-consistent* if $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \bar{\theta}$ as $n \to \infty$. If the convergence is almost sure w.r.t. $\mathbb{P}$, the estimator is called *strongly $\bar{\theta}$-consistent*.

Since $\widehat{\theta}_n$ takes its values in the "extended" space Θ_δ, one has to define a metric on this set in order that the previous convergence makes sense. In Θ it is of course the restriction of the usual metric on $\mathbb{R}^p$; and we will say that the distance between the extra point δ and any point in Θ equals 1. Therefore weak $\bar{\theta}$-consistency implies in particular that $\mathbb{P}(\widehat{\theta}_n = \delta) \to 0$, or otherwise formulated: $\widehat{\theta}_n$ solves the estimating equation (2.1) with a probability tending to one.

Theorem 2.5 *Under Condition 2.2 we can find a sequence $(\widehat{\theta}_n)$ of G_n-estimators which is weakly $\bar{\theta}$-consistent. Moreover this sequence is* eventually unique, *that is if $(\widehat{\theta}'_n)$ is any other weakly $\bar{\theta}$–consistent sequence of G_n–estimators, then $\mathbb{P}(\widehat{\theta}_n \neq \widehat{\theta}'_n) \to 0$ as $n \to \infty$.*

This theorem is proved in Section 6, and a similar proof gives the following result.

Theorem 2.6 *Suppose Condition 2.2 holds with almost sure convergence instead of convergence in probability. Then a sequence $(\widehat{\theta}_n)$ of G_n–estimators exists which is* strongly $\bar{\theta}$–consistent.

It is important to observe that so far we have made no assumption about identifiability. Consequently, although there is a sequence of G_n–estimators that is (weakly or strongly) $\bar{\theta}$–consistent, there might be other sequences that are not. For example, if G vanishes for another value $\bar{\theta}'$ and $\partial_\theta G(\bar{\theta}') \in \mathcal{M}_p^{\text{inv}}$, then there is another sequence of G_n–estimators that converges to $\bar{\theta}'$. An estimating function with such a property is obviously not of much practical use. Also, the reader can observe that the construction of $\widehat{\theta}_n$ in the proof of the previous theorems made use of the value $\bar{\theta}$. Thus this result is a mathematical existence

result, and the method can obviously not be used in statistical practice to choose a good G_n–estimator among several solutions to $G_n(\theta) = 0$ since $\bar{\theta}$ is unknown.

Thus the previous two theorems are powerful technical tools, but to obtain practically useful results, global properties of the estimating function are needed. Essentially, $\bar{\theta}$ must be the only root to $G(\theta) = 0$. Global uniqueness results can be obtained when Θ is compact and when Θ is homeomorphic to $\mathbb{R}^p$ (and hence is open). These two cases cover most practical situations. Recall that Θ is homeomorphic to $\mathbb{R}^d$ if there is a bijective bicontinuous mapping $\psi : \Theta \to \mathbb{R}^p$.

When Θ is homeomorphic to $\mathbb{R}^p$, to obtain a global uniqueness result, we must restrict attention to a class of estimators that are prevented from going to the boundary of Θ. We define such a class as follows. Pick an arbitrary point $\rho \in \Theta$, and define a *ρ–centered G_n–estimator* as any $\mathcal{F}_n$–measurable and Θ_δ-valued variable $\widehat{\theta}_n^\rho$ such that

$$\widehat{\theta}_n^\rho = \begin{cases} \operatorname{argmin} \left(d(\theta, \rho) : \theta \in \Theta, G_n(\theta) = 0\right) & \text{on the set } D_n \\ \delta & \text{on } (D_n)^c, \end{cases} \tag{2.6}$$

where $d(\theta, \theta') = |\psi(\theta) - \psi(\theta')|$ (ψ is the bijection $\Theta \to \mathbb{R}^p$), and δ and D_n are as in Definition 2.1. By the measurable selection theorem and the continuity of G_n, such estimators always exist.

Theorem 2.7 *1) If for some neighbourhood $M \subseteq \Theta$ of $\bar{\theta}$ we have both (2.3) and that for all $\varepsilon > 0$, $\mathbb{P}\left(\inf_{M, |\theta - \bar{\theta}| > \varepsilon} |G(\theta)| > 0\right) = 1$, then any sequence $\widehat{\theta}_n$ of G_n–estimators satisfies*

$$\mathbb{P}(\widehat{\theta}_n \in M, |\widehat{\theta}_n - \bar{\theta}| > \varepsilon) \to 0 \quad \textit{for all } \varepsilon > 0. \tag{2.7}$$

2) Assume that Condition 2.2 holds for all compact subsets $M \subseteq \Theta$, and that $\bar{\theta} \in \operatorname{int} \Theta$ is the unique root of the equation $G(\theta) = 0$. Then the conditions in 1) hold for any compact neighbourhood of $\bar{\theta}$.

(a) Suppose Θ is compact. Then any sequence of G_n–estimators is weakly $\bar{\theta}$–consistent, and this sequence is eventually unique.

(b) Suppose Θ is homeomorphic to $\mathbb{R}^p$, and choose any $\rho \in \Theta$. Then any sequence of ρ–centered G_n–estimators is weakly $\bar{\theta}$–consistent, and this sequence is eventually unique, in the sense that if $\widehat{\theta}_n^\rho$ and $\widehat{\theta}_n'^{\rho'}$ are two sequences of, respectively, ρ and ρ'–centered G_n–estimators, then $\mathbb{P}(\widehat{\theta}_n^\rho \neq \widehat{\theta}_n'^{\rho'}) \to 1$.

Note that the asymptotic behavior of $\widehat{\theta}_n^\rho$ in (b) does not depend on the choice of ρ.

2.2 Rate of convergence and asymptotic distribution

The results in the previous subsection ensure only the existence of a solution converging to $\bar{\theta}$, but say nothing about the rate of the convergence. To obtain results about the rate, we need a stronger condition like, for instance, the following.

Condition 2.8 *There exists a sequence of positive real numbers a_n increasing to infinity, such that the sequence of random variables $a_n G_n(\bar{\theta})$ is stochastically bounded, i.e. such that for every $\varepsilon > 0$ there exists a $K > 0$ such that $\mathbb{P}\left(|a_n G_n(\bar{\theta})| > K\right) < \varepsilon$ for all n.*

If the sequence $a_n G_n(\bar{\theta})$ converges in distribution, it is stochastically bounded. The sequence a_n can obviously be chosen in many ways. As appears from the following theorem, the most interesting choice is a sequence that goes to infinity as fast as possible.

Theorem 2.9 *Under Conditions 2.2 and 2.8 there is a sequence $(\widehat{\theta}_n)$ of G_n–estimators such that*

$$\lim_{c\to\infty} \sup_n \mathbb{P}\left(|a_n(\hat{\theta}_n - \bar{\theta})| \leq c\right) = 1 \tag{2.8}$$

or in other words the sequence $(a_n|\widehat{\theta}_n - \bar{\theta}|)$ is stochastically bounded. Moreover, any sequence $(\widehat{\theta}_n)$ of G_n–estimators which is weakly $\bar{\theta}$–consistent satisfies (2.8).

In some cases the coordinates of the estimator $\hat{\theta}_n$ do not all converge to $\bar{\theta}$ at the same rate, see Section 5. When this happens, rates of convergence follow from the result below on the asymptotic distribution of G_n–estimators, which can usually be derived from the asymptotic distribution of the functions G_n and their derivatives, as stated in the next condition.

In that condition we have a p-dimensional variable Z, defined on an extension of the space $(\Omega, \mathcal{F}, \mathbb{P})$. By an extension we mean a triple $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ with $\bar{\Omega} = \Omega \times \Omega'$ and $\bar{\mathcal{F}} = \mathcal{F} \otimes \mathcal{F}'$ with $(\Omega', \mathcal{F}')$ another measurable space, and $\bar{\mathbb{P}}$ is a probability measure satisfying $\bar{\mathbb{P}}(A \times \Omega') = \mathbb{P}(A)$ for all $A \in \mathcal{F}$. Any variable on Ω or Ω' is extended in a trivial way as a variable on $\bar{\Omega}$. Then we say that a sequence Z_n of p-dimensional random variables on Ω converges stably in law to Z if $\mathbb{E}(Yh(Z_n)) \to \bar{\mathbb{E}}(Yh(Z))$ for any bounded variable Y on $(\Omega, \mathcal{F})$ and any continuous bounded function h on $\mathbb{R}^p$, and we write $Z_n \xrightarrow{\mathcal{L}_{\rm st}} Z$. This automatically implies $Z_n \xrightarrow{\mathcal{L}} Z$.

Condition 2.10 *There exist a sequence $A_n \in \mathcal{M}_p^{\rm inv}$ with each entry of A_n^{-1} tending to zero, a random vector Z on an extension of the space, and a random $\mathcal{M}_p$–valued function W, with $W(\bar{\theta})$ almost surely in $\mathcal{M}_p^{\rm inv}$, such that for a neighbourhood M of $\bar{\theta}$ we have the following two properties:*

$$A_n G_n(\bar{\theta}) \xrightarrow{\mathcal{L}_{\rm st}} Z, \tag{2.9}$$

$$\sup_{\theta \in M} \| A_n \partial_\theta G_n(\theta) A_n^{-1} - W(\theta)\| \xrightarrow{\mathbb{P}} 0. \tag{2.10}$$

In particular, this condition implies that

$$\begin{pmatrix} A_n G_n(\bar{\theta}) \\ A_n\, \partial_\theta G_n(\bar{\theta})\, A_n^{-1} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} Z \\ W(\bar{\theta}) \end{pmatrix}. \tag{2.11}$$

When $W(\bar{\theta})$ is non-random, this is implied by (2.10), plus $A_n G_n(\bar{\theta}) \xrightarrow{\mathcal{L}} Z$ instead of (2.9).

Theorem 2.11 *Assume Conditions 2.2 and 2.10 holds, and let $\widehat{\theta}_n$ be a weakly $\bar{\theta}$–consistent sequence of G_n–estimators. Then*

$$A_n(\widehat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{L}_{st}} -W(\bar{\theta})^{-1} Z \tag{2.12}$$

and

$$A_n \partial_\theta G_n(\hat{\theta}_n)(\widehat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{L}_{st}} -Z. \tag{2.13}$$

If the sequence A_n^{-1} does not go fast enough to 0, Condition 2.10 may hold with $Z = 0$, and (2.12) only gives a rate of convergence which is not sharp. So, this result becomes really interesting when Condition 2.10 holds with a variable Z that is *strongly non-degenerate*, in the sense that the support of its law is not included in any proper linear subspace of $\mathbb{R}^p$.

Quite often Z is, conditionally on $\mathcal{F}$, centered Gaussian with an invertible covariance matrix $V = V(\omega)$. If V is non-random this amounts to having Z independent of $\mathcal{F}$ and Gaussian. In the general conditionally Gaussian case, the limit distribution in (2.12) is the normal

variance-mixture with characteristic function $s \mapsto \mathbb{E}\left(\exp\left[-\frac{1}{2}s^* W(\bar{\theta})^{-1} V W(\bar{\theta})^{*-1} s\right]\right)$, which is Gaussian if both V and $W(\bar{\theta})$ are non-random. If one can construct weakly consistent estimators $\widehat{V}_n$ for V, in the sense that $\widehat{V_n} \xrightarrow{\mathbb{P}} V$ and each $\widehat{V}_n$ is positive definite, then (2.13) implies that (I_p is the $p \times p$ identity matrix):

$$\widehat{V}_n^{-1/2} A_n \, \partial_\theta G_n(\hat{\theta}_n) \, (\widehat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, I_p),$$

from which one can easily find confidence regions for the parameter $\bar{\theta}$. This may be difficult using the non-standard distribution in (2.12).

3 Ergodic processes

In this subsection we consider the case where the observed process is ergodic and present simple conditions that imply the previous general assumptions. We assume that we have a sequence of random variables $X_1, X_2, \ldots$ with a measurable state space $(D, \mathcal{D})$, which is ergodic under the true measure $\mathbb{P}$. By ergodic we here mean that, for every integer $r \geq 1$, there is a probability measure Q_r on D^r, such that for any function $f : D^r \mapsto \mathbb{R}$ that is Q_r–integrable,

$$\frac{1}{n} \sum_{i=r}^{n} f(X_{i-r+1}, \ldots, X_i) \xrightarrow{\mathbb{P}} Q_r(f), \tag{3.1}$$

where $Q_r(f)$ denotes the integral of f with respect to Q_r. It is a weak form of ergodicity, which encompasses the case where we observe a continuous time Markov process Y at equidistant time points, i.e. $X_n = Y_{n\Delta}$, $\Delta > 0$. Suppose the state space of Y is a domain $D \subseteq \mathbb{R}^d$, and the transition kernels of Y have positive Lebesgue-densities $p_t(x, y)$, so $\mathbb{P}(Y_t \in A \mid Y_0 = x) = \int_A p_t(x, y)\, dy$ for all $t > 0$ and $x \in D$. If Y has a unique invariant probability measure μ, and if $p_t(x, y)$ is a continuous function of x for all $t > 0$ and $y \in D$, then (3.1) holds for any initial measure η on D. In this case, as an example, $Q_2(dx, dy) = p_\Delta(x, y)\mu(dx)dy$.

We assume that at stage n we observe the n first variables $X_1, \ldots, X_n$, and we consider estimating functions of the form

$$G_n(\theta) = \frac{1}{n} \sum_{i=r}^{n} g(X_{i-r+1}, \ldots, X_i; \theta), \tag{3.2}$$

where $g : D^r \times \Theta \mapsto \mathbb{R}^p$ is jointly measurable and satisfies the following assumption.

Condition 3.1 *There is parameter value $\bar{\theta} \in \operatorname{int}\Theta$ and a neighbourhood N of $\bar{\theta}$ in Θ, such that:*
(1) The function $g(\theta) : (x_1, \ldots, x_r) \mapsto g(x_1, \ldots, x_r; \theta)$ is Q_r–integrable for all $\theta \in N$, and $Q_r\left(g(\bar{\theta})\right) = 0$.
(2) The function $\theta \mapsto g(x_1, \ldots, x_r; \theta)$ is C^1 on N for all $(x_1, \ldots, x_r)$ in D^r.
(3) For all compact subsets M of N there is a Q_r–integrable function $\bar{g}_M$ on D^r such that $|\partial_\theta g(\theta)| \leq \bar{g}_M$ for all $\theta \in M$.
(4) The $p \times p$ matrix $Q_r(\partial_\theta g(\bar{\theta}))$ is invertible.

Note that Condition 3.1 (3) is the property that the function $\partial_\theta g(\theta)$ is locally dominated Q_r–integrable for $\theta \in N$. This is a traditional condition in the statistical literature.

Theorem 3.2 *Under Condition 3.1, the estimating functions G_n satisfy Condition 2.2 for any compact subset $M \subseteq N$ and the (non-random) $\mathbb{R}^p$-valued function $G(\theta) = Q_r(g(\theta))$. Hence there exists an eventually unique sequence of weakly $\bar{\theta}$–consistent G_n–estimators.*

If further $N = \Theta$ and $\bar{\theta}$ is the unique root of the function $G(\theta)$, then any sequence $\widehat{\theta}_n$ of G_n–estimators satisfies (2.7) for all compact neighbourhoods M of $\bar{\theta}$. Moreover, if Θ is compact, then any sequence $\widehat{\theta}_n$ of G_n–estimators is weakly $\bar{\theta}$-consistent, and if Θ is homeomorphic to $\mathbb{R}^p$, then for any $\rho \in \Theta$, any sequence $\widehat{\theta}_n^\rho$ of ρ-centered G_n–estimators is weakly $\bar{\theta}$-consistent.

Suppose that we assumed ergodicity in a slightly stronger sense, namely that in (3.1) the convergence takes place $\mathbb{P}$–almost surely. This is, for instance, the case for a discretely observed continuous time Markov process that satisfies the conditions given above. Then in Condition 2.2 we obtain almost sure convergence as well, and by Theorem 2.6 we see that there exists a sequence of strongly $\bar{\theta}$–consistent G_n–estimators.

Finally, let us assume that the estimating functions G_n also satisfy a central limit theorem:

$$\sqrt{n}\, G_n(\bar{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=r}^{n} g(X_{i-r+1}, \ldots, X_i; \bar{\theta}) \xrightarrow{\mathcal{L}} N(0, V(\bar{\theta})) \tag{3.3}$$

for some (necessarily non–random) $p \times p$–matrix $V(\bar{\theta})$. Then it follows from Theorem 2.11 that any sequence $\widehat{\theta}_n$ of weakly $\bar{\theta}$–consistent G_n–estimators satisfies

$$\sqrt{n}\,(\widehat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Q_r(\partial_\theta g(\bar{\theta}))^{-1}\, V(\bar{\theta})\, Q_r(\partial_\theta g(\bar{\theta}))^{*-1}).$$

4 Longitudinal data

The previous section applies, in particular, when the observations X_i are i.i.d., the situation for which estimating functions were initially introduced. In this case one naturally takes $r = 1$, (3.1) is the usual LLN (with a.s. convergence) and (3.3) the usual CLT. However, each observation X_i may have a complex structure. An interesting example is longitudinal data, where each X_i consists of observations of a stochastic process. This kind of data, that are also referred to as panel data, have received a lot of attention in the statistical and econometric literature. In this branch of statistics, estimating functions have frequently been applied. A classical text is Diggle *et al.* (2002). For applications of diffusion processes in a longitudinal data context, see Pedersen (2000) and Bibbona and Ditlevsen (2012).

As an example, we consider the case where the X_i's are n independent copies of a discretely observed stationary Markov process Y. More specifically, we have a state space D, and for each θ a stationary Markov process Y^θ with transition semi-group $(P_t^\theta)_{t\geq 0}$ and stationary initial distribution μ^θ. The ith longitudinal observation is $X_i = (Y^i_{j\Delta} : j = 0, 1, \ldots, m)$, where $\Delta > 0$, m is a fixed integer, and $Y^1, \ldots, Y^n$ are i.i.d. copies of Y^{θ_0}, where θ_0 is the true parameter value.

We consider the estimating function

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} h(Y^i_{(j-1)\Delta}, Y^i_{j\Delta}; \theta),$$

where h is a measurable function on $D^2 \times \Theta$ which satisfies

$$\int_D h(x_1, x_2; \theta)\, P_\Delta^\theta(x_1, dx_2) = 0 \tag{4.1}$$

for all $\theta \in \Theta$ and all $x_1 \in D$. This means that the inner sum defines a *martingale estimating function* for the ith Markov process. In contrast to the previous section, m is here fixed, and we exploit the martingale property only to compute the limiting variance. Below, $Q(dx_1, dx_2) = \mu^{\theta_0}(dx_1)\, P^{\theta_0}(x_1, dx_2)$ is the law of (Y_0^1, Y_Δ^1) under the true measure. Besides (4.1), we impose the following condition on h:

Condition 4.1
(1) The function $h(\theta) :\ (x_1, x_2) \mapsto h(x_1, x_2; \theta)$ is square–integrable with respect to Q for all $\theta \in \Theta$.
(2) The function $\theta \mapsto h(x_1, x_2; \theta)$ is C^1 for all $(x_1, x_2) \in D^2$.
(3) For all compact subsets M of Θ there is a Q–integrable function $\bar{h}_M$ on D^2 such that $|\partial_\theta h(\theta)| \le \bar{h}_M$ for all $\theta \in M$.
(4) The $p \times p$ matrix $Q(\partial_\theta h(\theta_0))$ is invertible.

This seems identical to Condition 3.1 with $r = 2$ and $\bar{\theta} = \theta_0$ ($Q(h(\theta_0)) = 0$ follows from (4.1)), but the meaning of Q here is quite different from the meaning of Q_2 in that condition.

Theorem 4.2 *Under Condition 4.1 there exists a sequence $\widehat{\theta}_n$ of strongly consistent G_n–estimators, eventually unique and satisfying (as $n \to \infty$)*

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_0) \ \xrightarrow{\mathcal{L}}\ \mathcal{N}_p(0, m^{-1}\, V(\theta_0)), \tag{4.2}$$

where $V(\theta_0) = (Q(\partial_\theta h(\theta_0)))^{-1}\, Q(h(\theta_0)\, h(\theta_0)^*)\, (Q(\partial_\theta h(\theta_0)))^{\star -1}$.

If further $Q(g(\theta)) \ne 0$ for all $\theta \ne \theta_0$, then any sequence $\widehat{\theta}_n$ of G_n–estimators satisfies (2.7) with $\bar{\theta} = \theta_0$ for all compact neighbourhoods K of θ_0. Moreover, if Θ is compact, then any sequence $\widehat{\theta}_n$ of G_n–estimators satisfies (4.2), and if Θ is homeomorphic to $\mathbb{R}^p$, then for any $\rho \in \Theta$ and any sequence $\widehat{\theta}_n^\rho$ of ρ–centered G_n–estimators satisfies (4.2).

5 High frequency observation of a diffusion

In this section we present examples which illustrate that it is necessary to allow the limiting functions in Conditions 2.2 and 2.10 to be random, and that it is necessary to allow different rates of convergence for the coordinates of the estimators.

For each value $\theta = (\alpha, \beta)$ in a subset $\Theta = A \times B$ of $\mathbb{R}^2$, we consider the one–dimensional diffusion process given by the stochastic differential equation

$$dX_t^\theta = a(X_t^\theta, \alpha)\, dt + b(X_t^\theta; \beta)\, dW_t, \qquad X_0^\theta = x_0, \tag{5.1}$$

with W a standard Wiener process on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \ge 0}, \mathbb{P})$, where $\mathcal{F}$ is the $\mathbb{P}$–completion of $\bigvee_{t>0} \mathcal{F}_t$ and $(\mathcal{F}_t)$ the filtration generated by W.

We observe $X = X^{\theta_0}$ at the times $i\Delta_n$, $i = 0, 1, \ldots, n$, for the true parameter value $\theta_0 = (\alpha_0, \beta_0)$, and we assume that $\Delta_n \to 0$ as $n \to \infty$. Thus we have observation in the time interval $[0, T_n]$ only, where the total time span is $T_n = n\Delta_n$. Since X_0 is observed, it is no restriction to assume that the starting point is a non-random number x_0, independent of θ.

We make the following smoothness assumptions on the coefficients. These could be substantially weakened, at the price of more complex proofs. That the state space is the entire set $\mathbb{R}$ could also be relaxed.

Assumption 5.1
(1) The function b is C^3 on $\mathbb{R} \times B$, and all its derivatives are of polynomial growth in x, uniformly in $\beta \in K$ for all compact subsets $K \subseteq B$.
(2) The function a is C^2 on $\mathbb{R} \times A$, and all its derivatives are of polynomial growth in x, uniformly in $\alpha \in K$ for all compact subsets $K \subseteq A$.
(3) The functions $a(\cdot;\alpha)$ and $b(\cdot;\beta)$ are globally Lipschitz for all $\theta = (\alpha, \beta)$.
(4) $\inf_{x\in\mathbb{R},\,\beta\in K} b(x;\beta) > 0$ for all compact subsets $K \subseteq B$.

This implies in particular that (5.1) has a unique strong solution, which is Markov, and also that for any $m \geq 0$, $t \geq 0$ and any compact subset $K \subseteq \Theta$:

$$\sup_{\theta\in K,\,0\leq\Delta\leq 1} \Delta^{-m/2}\,\mathbb{E}(\,\sup_{s\in[0,\Delta]}\,|X^\theta_{t+s} - X^\theta_t|^m \mid \mathcal{F}_t) < \infty. \tag{5.2}$$

The situation is quite different when $T_n \equiv T$ for some fixed $T > 0$ and when $T_n \to \infty$ as $n \to \infty$. Below, we consider these two cases.

5.1 Fixed time span

When $T_n \equiv T$, it is well known that the drift cannot be estimated consistently, so we consider only estimation of the second component β of the parameter θ. In other words, the set A consists of a single point, or $a(x,\alpha) = a(x)$ does not depend on a parameter. The drift coefficient still satisfies the relevant conditions in Assumption 5.1, but it may be unknown (the estimating functions (5.3) does not depend on a), so we are solving a semi-parametric problem.

With the notation $c(x,\beta) = b(x,\beta)^2$ and $\Delta^n_i X = X_{i\Delta_n} - X_{(i-1)\Delta_n}$, a simple estimating function for β, which yields the estimator proposed by Genon-Catalot and Jacod (1993), is given by

$$G_n(\beta) = \sum_{i=1}^n \frac{\partial_\beta c(X_{(i-1)\Delta_n};\beta)}{c(X_{(i-1)\Delta_n};\beta)^2}\Big((\Delta^n_i X)^2 - \Delta_n\, c(X_{(i-1)\Delta_n};\beta)\Big). \tag{5.3}$$

Theorem 5.2 *Suppose $T_n \equiv T$ and that Assumption 5.1 and the identifiability condition $\int_0^T |\partial_\theta b(X_t,\beta_0)|\,dt > 0$ outside a $\mathbb{P}$-null set hold. Then the estimating function G_n satisfies the Conditions 2.2 and 2.10 with $\bar\theta = \theta_0 = \beta_0$, any compact subset $M \subseteq B$ and the random functions*

$$G(\beta) = \int_0^T \frac{\partial_\beta c(X_t;\beta)}{c(X_t;\beta)^2}(c(X_t;\beta_0) - c(X_t;\beta))\,dt, \tag{5.4}$$

$$W(\beta) = \partial_\beta G(\beta) = \int_0^T \left[\partial_\beta\Big(\frac{\partial_\beta c(X_t;\beta)}{c(X_t;\beta)^2}\Big)(c(X_t;\beta_0) - c(X_t;\beta)) - \frac{(\partial_\beta c(X_t;\beta))^2}{c(X_t;\beta)^2}\right] dt. \tag{5.5}$$

Moreover, $A_n = \sqrt{n}$ and Z is a random variable which conditionally on $\mathcal{F}$ is centered Gaussian with variance $-2T\,W(\beta_0)$.

Suppose further that $G(\beta) \neq 0$ for all $\beta \neq \beta_0$, then any sequence $\widehat\beta_n$ of G_n–estimators satisfies (2.7) with $\widehat\theta_n = \widehat\beta_n$ and $\bar\theta = \beta_0$ for all compact neighbourhoods K of β_0. Moreover, if B is compact, then any sequence $\widehat\beta_n$ of G_n–estimators satisfies (2.12) and (2.13), and if B is an open (finite or infinite) interval, then for all $\rho \in B$, any sequence $\widehat\beta^\rho_n$ of ρ–centered G_n–estimators satisfies (2.12) and (2.13).

The estimator studied here is efficient and has the optimal rate of convergence, see Gobet (2001). A general theory for approximate martingale estimating functions when $T_n \equiv T$ can be found in Jakobsen and Sørensen (2017). Another general class of estimators was investigated in Genon-Catalot and Jacod (1993).

5.2 Time span going to infinity

Now, we assume that $T_n \to \infty$, so drift parameters can be consistently estimated, and we consider the full parameter space $\Theta = A \times B$. We need the following ergodicity assumption.

Assumption 5.3 *Under the true parameter value θ_0 the solution of (5.1) is ergodic, in the sense that there is a probability measure μ_{θ_0} on $\mathbb{R}$ (necessarily an invariant measure for the Markov process X^{θ_0}) such that any μ_{θ_0}–integrable function f satisfies*

$$\frac{1}{n}\sum_{i=0}^{n} f(X^{\theta_0}_{i\Delta_n}) \xrightarrow{\mathbb{P}} \mu_{\theta_0}(f). \tag{5.6}$$

Moreover, $\sup_{t>0} \mathbb{E}(|X^{\theta_0}_t|^m) < \infty$ for all $m \geq 0$, and hence $\int |x|^m \, \mu_{\theta_0}(dx) < \infty$ as well.

Let us also state an identifiability assumption.

Assumption 5.4 *We have $\int_{\mathbb{R}} |\partial_\beta b(x, \beta_0)| \, \mu_{\theta_0}(dx) > 0$ and $\int_{\mathbb{R}} |\partial_\alpha a(x, \alpha_0)| \, \mu_{\theta_0}(dx) > 0$.*

We use the two-dimensional estimation function $G_n = (G_n^1, G_n^2)$ given for $\theta = (\alpha, \beta)$ by

$$\begin{aligned} G_n^1(\theta) &= \tfrac{1}{T_n} \sum_{i=1}^{n} \tfrac{\partial_\alpha a(X_{(i-1)\Delta_n};\alpha)}{c(X_{(i-1)\Delta_n};\beta)} \Big(\Delta_i^n X - \Delta_n \, a(X_{(i-1)\Delta_n}; \alpha) \Big) \\ G_n^2(\theta) &= \tfrac{1}{T_n} \sum_{i=1}^{n} \tfrac{\partial_\beta c(X_{(i-1)\Delta_n};\beta)}{c(X_{(i-1)\Delta_n};\beta)^2} \Big((\Delta_i^n X)^2 - \Delta_n \, c(X_{(i-1)\Delta_n}; \beta) \Big). \end{aligned} \tag{5.7}$$

Note that G_n^2 equals the estimating functions (5.3).

Theorem 5.5 *Suppose $T_n \to \infty$ and $n\Delta_n^2 \to 0$, and that Assumptions 5.1, 5.3 and 5.4 hold. Then the estimating function G_n given by (5.7) satisfies Conditions 2.2 and 2.10 for any compact subset $M \subseteq \Theta$ and with $\bar{\theta} = \theta_0$, $A_n = \operatorname{diag}(\sqrt{T_n}, \sqrt{n})$, and the non-random functions*

$$\begin{aligned} G^1(\theta) &= \int_{\mathbb{R}} \frac{\partial_\alpha a(x;\alpha)}{c(x;\beta)} (a(x;\alpha_0) - a(x;\alpha)) \, \mu_{\theta_0}(dx), \\ G^2(\theta) &= \int_{\mathbb{R}} \frac{\partial_\beta c(x;\beta)}{c(x;\beta)^2} (c(x;\beta_0) - c(x;\beta)) \, \mu_{\theta_0}(dx) \end{aligned}$$

$$\begin{aligned} W^{11}(\theta) &= \partial_\alpha G^1(\theta) = \int_{\mathbb{R}} \left[\partial_\alpha \Big(\frac{\partial_\alpha a(x;\alpha)}{c(x;\beta)} \Big) (a(x;\alpha_0) - a(x;\alpha)) - \frac{(\partial_\alpha a(x;\alpha))^2}{c(x;\beta)} \right] \mu_{\theta_0}(dx) \\ W^{22}(\theta) &= \partial_\beta G^2(\theta) = \int_{\mathbb{R}} \left[\partial_\beta \Big(\frac{\partial_\beta c(x;\beta)}{c(x;\beta)^2} \Big) (c(x;\beta_0) - c(x;\beta)) - \frac{(\partial_\beta c(x;\beta))^2}{c(x;\beta)^2} \right] \mu_{\theta_0}(dx) \\ W^{12}(\theta) &= \partial_\beta G^1(\theta) = - \int_{\mathbb{R}} \frac{\partial_\alpha a(x;\alpha) \, \partial_\beta c(x;\beta)}{c(x;\beta)^2} \, (a(x;\alpha_0) - a(x;\alpha)) \, \mu_{\theta_0}(dx) \\ W^{21}(\theta) &= \partial_\alpha G^2(\theta) = 0, \end{aligned}$$

and with $Z = (Z^1, Z^2)$ a two-dimensional variable independent of $\mathcal{F}$, where Z^1 and Z^2 are independent centered Gaussian with variances $-W^{11}(\theta_0)$ and $-W^{22}(\theta_0)$.

Suppose further that $G(\theta) \neq 0$ for all $\theta \neq \theta_0$, then any sequence $\widehat{\theta}_n$ of G_n–estimators satisfies (2.7) with $\bar{\theta} = \theta_0$ for all compact neighbourhoods K of θ_0. Moreover, if Θ is compact, then any sequence $\widehat{\theta}_n$ of G_n–estimators satisfies (2.12) and (2.13), and if Θ is homeomorphic to $\mathbb{R}^2$, then for all $\rho \in \Theta$, any sequence $\widehat{\theta}_n^{\rho}$ of ρ–centered G_n–estimators satisfies (2.12) and (2.13).

The estimator studied here is efficient and has an optimal rate of convergence, see Gobet (2002). A general theory for approximate martingale estimating functions for diffusion processes under the asymptotic scenario considered in this subsection can be found in Sørensen (2017).

6 Proofs

The proof of Theorem 2.5 is based on the fixed point theorem for contractions. A mapping f from a subset M of $\mathbb{R}^p$ into $\mathbb{R}^p$ is called a *contraction* if there exists a constant $C \in (0, 1)$, called a contraction constant, such that $|f(x) - f(y)| \leq C|x - y|$ for all $x, y \in M$. A proof of the following lemma can for instance be found at page 229 in Loomis and Sternberg (1968). Below, $\bar{B}_r(x)$ is the closed ball of $\mathbb{R}^p$ with radius r and center x.

Lemma 6.1 *Let $f : \bar{B}_r(x_0) \mapsto \mathbb{R}^p$ be a contraction such that $|f(x_0) - x_0| \leq (1 - C)r$, where C is the contraction constant. Then f has a unique fixed point x (i.e., $f(x) = x$) in $\bar{B}_r(x_0)$, which is the limit of the sequence x_n started at x_0 and defined by induction through $x_{n+1} = f(x_n)$.*

Specifically, the fixed point theorem is used to prove the following lemma.

Lemma 6.2 *Let f be a differentiable mapping from a closed subset M of $\mathbb{R}^p$ into $\mathbb{R}^p$, and let $A \in \mathcal{M}_p^{\text{inv}}$. Define $\lambda = \frac{1}{2}\|A^{-1}\|^{-1}$. If*

$$\|\partial_x f(x) - A\| \leq \lambda$$

on a ball $\bar{B}_r(x_0) \subseteq M$, then any point y in $\bar{B}_{\lambda r}(f(x_0))$ is the image $y = f(x)$ of a unique point x in $\bar{B}_r(x_0)$.

Proof. Choose $y \in \bar{B}_{\lambda r}(f(x_0))$ and define the function $\phi(x) = x + A^{-1}(y - f(x))$. It is sufficient to prove that ϕ has a unique fixed point in $\bar{B}_r(x_0)$. Since $\partial_x \phi(x) = I - A^{-1}\partial_x f(x)$ for all $x \in \bar{B}_r(x_0)$, it follows that

$$\|\partial_x \phi(x)\| \leq \|A^{-1}\|\|A - \partial_x f(x)\| \leq \tfrac{1}{2}.$$

Thus ϕ is a contraction on $\bar{B}_r(x_0)$ with contraction constant $1/2$. Since

$$|\phi(x_0) - x_0| = |A^{-1}(y - f(x_0))| \leq \|A^{-1}\|\lambda r = r/2,$$

the result follows from the previous lemma. □

Proof of Theorem 2.5. 1) Let us introduce the random variables

$$\Lambda = \tfrac{1}{2}\|\partial_\theta G(\bar{\theta})^{-1}\|^{-1},$$

$$Y(\varepsilon) = \sup_{\theta:|\theta-\bar\theta|\le\varepsilon} \|\partial_\theta G(\theta) - \partial_\theta G(\bar\theta)\|,$$
$$Z_n = \sup_{\theta\in M} \|\partial_\theta G_n(\theta) - \partial_\theta G(\theta)\|,$$

which are $\mathcal{F}$–measurable (because $\mathcal{F}$ is $\mathbb{P}$-complete), and also the $\mathcal{F}$–measurable sets

$$C_{n,\varepsilon} = \left\{Y(\varepsilon) \le \tfrac{1}{2}\Lambda\right\} \cap \left\{Z_n \le \tfrac{1}{2}\Lambda\right\} \cap \left\{|G_n(\bar\theta)| \le \Lambda\varepsilon\right\}.$$

On the set $C_{n,\varepsilon}$ we have $Y(\varepsilon) + Z_n \le \Lambda$, hence $\|\partial_\theta G_n(\theta) - \partial_\theta G(\bar\theta)\| \le \Lambda$ whenever $|\theta - \bar\theta| \le \varepsilon$, and also $|G_n(\bar\theta)| \le \Lambda\varepsilon$. Thus for any given ω in $C_{n,\varepsilon}$ we can apply Lemma 6.2 with $f = G_n$ and $\lambda = \Lambda$ and $A = \partial_\theta G(\bar\theta)$ and $r = \varepsilon$, to get

$$\omega \in C_{n,\varepsilon} \;\Rightarrow\; \text{there is a unique } \theta_{n,\varepsilon}(\omega) \in \bar B_\varepsilon(\bar\theta) \text{ with } G_n(\theta_{n,\varepsilon}(\omega),\omega) = 0. \tag{6.1}$$

Moreover by Lemma 6.1 and the proof of Lemma 6.2, for each $\omega \in C_{n,\omega}$ we have $\theta_{n,\varepsilon} = \lim_p z_p$ for the sequence defined inductively by $z_0 = \bar\theta$ and $z_{p+1} = z_p - \partial_\theta G(\bar\theta) G_n(z_p)$. Hence $\theta_{n,\omega}$ (restricted to $C_{n,\varepsilon}$) is $\mathcal{F}_n$–measurable.

2) Let us now prove the existence of a sequence $\varepsilon_n \downarrow 0$ which satisfies

$$\mathbb{P}(C_{n,\varepsilon_n}) \;\to\; 1. \tag{6.2}$$

For this, we first recall the well known fact that a sequence of real–valued variables V_n goes to 0 in probability if and only if there is a sequence $\varepsilon_n \downarrow 0$ such that $\mathbb{P}(|V_n| \ge \varepsilon_n) \to 0$.

Condition 2.2 yields $\mathbb{P}(\Lambda > 0) = 1$ and $G_n(\bar\theta) \stackrel{\mathbb{P}}{\to} 0$ and $Z_n \stackrel{\mathbb{P}}{\to} 0$ and $\limsup_{\varepsilon\to 0} Y(\varepsilon) = 0$ almost surely. Hence $G_n(\bar\theta)/\Lambda \stackrel{\mathbb{P}}{\to} 0$ and $Z_n/\Lambda \stackrel{\mathbb{P}}{\to} 0$ and $\limsup_{\varepsilon\to 0} Y(\varepsilon)/\Lambda = 0$ almost surely, and we deduce that, for some sequence $\varepsilon_n \downarrow 0$,

$$\mathbb{P}(C_{n,\varepsilon_n}) \ge 1 - \mathbb{P}\left(Y(\varepsilon_n)/\Lambda > \tfrac{1}{2}\right) - \mathbb{P}\left(Z_n/\Lambda > \tfrac{1}{2}\right) - \mathbb{P}\left(|G_n(\bar\theta)|/\Lambda > \varepsilon_n\right) \;\to\; 1, \tag{6.3}$$

which yields (6.2).

3) Now we are ready to prove the existence of a weakly $\bar\theta$–consistent sequence $\widehat\theta_n$ of G_n–estimators. First, we choose arbitrary G_n–estimators $\widehat\theta'_n$, which are known to exist. Then, with ε_n as above, we define $\widehat\theta_n$ to be equal to θ_{n,ε_n} on the set C_{n,ε_n} and to $\widehat\theta'_n$ on the complement of this set. This gives us an $\mathcal{F}_n$-measurable variable $\widehat\theta_n$, whose weak $\bar\theta$–consistency readily follows from the fact that $|\theta_{n,\varepsilon_n} - \bar\theta| \le \varepsilon_n$, plus $\varepsilon_n \to 0$ and (6.2).

4) It remains to prove the last claim. We assume that we have two sequences $(\widehat\theta_n)$ and $(\widehat\theta'_n)$ of G_n–estimators, both of them weakly $\bar\theta$-consistent, and we want to prove $\mathbb{P}(\widehat\theta_n \ne \widehat\theta'_n) \to 0$. Since $\widehat\theta_n - \bar\theta$ and $\widehat\theta'_n - \bar\theta$ go to 0 in probability, we can find a sequence $\varepsilon_n \downarrow 0$ such that (6.2) holds, together with

$$\mathbb{P}(|\widehat\theta_n - \bar\theta| \ge \varepsilon_n) \;\to\; 0, \qquad \mathbb{P}(|\widehat\theta'_n - \bar\theta| \ge \varepsilon_n) \;\to\; 0. \tag{6.4}$$

Since both $\widehat\theta_n$ and $\widehat\theta'_n$ solve the estimating equation $G_n(\theta) = 0$ when this equation has a solution, we readily deduce from (6.1) that on the set C_{n,ε_n} and if $|\widehat\theta_n - \bar\theta| \le \varepsilon_n$ and $|\widehat\theta'_n - \bar\theta| \le \varepsilon_n$, then necessarily $\widehat\theta_n = \widehat\theta'_n$. Hence $\mathbb{P}(\widehat\theta_n \ne \widehat\theta'_n) \to 0$ follows from (6.2) and (6.4). □

Proof of Theorem 2.7. By d we denote the Euclidean distance, except in case 2 (b) where d denotes the distance used in (2.6). In both cases $d(\theta,\delta) = 1$ for all $\theta \in \Theta$ by convention. For

any neighbourhood $M \subseteq \Theta$ of $\bar\theta$ and $\varepsilon \in (0,1)$, we define $Z(M)_n = \sup_{\theta\in M} |G_n(\theta)-G(\theta)|$ and $Y(M,\varepsilon) = \inf_{\theta\in M,\, d(\theta,\bar\theta)>\varepsilon} |G(\theta)|$. Under the assumptions in 1), $Y(M,\varepsilon) > 0$ and $Z(M)_n \stackrel{\mathbb{P}}{\longrightarrow} 0$, so as $n \to \infty$:

$$\mathbb{P}(Z(M)_n < Y(M,\varepsilon)) \;\to\; 1. \tag{6.5}$$

Because $\{\widehat\theta_n \in M\} \subseteq \{Z(M)_n \geq |G(\widehat\theta_n)|\}$, then any sequence $\widehat\theta_n$ of G_n–estimators satisfies

$$d(\widehat\theta_n, \bar\theta) \leq \varepsilon \;\text{ on the set }\; \{\widehat\theta_n \in M\} \cap \{Z(M)_n < Y(M,\varepsilon)\}. \tag{6.6}$$

This implies the conclusion in 1).

The conditions in 2) obviously implies that the conditions in 1) hold for any compact neighbourhood of $\bar\theta$. In case (a), (6.5) and (6.6) with $M = \Theta$, plus the facts that $\{\widehat\theta_n \in \Theta\} = D_n$ and $\mathbb{P}(D_n) \to 1$ by Theorem 2.5, yield the weak $\bar\theta$-consistency of $\widehat\theta_n$. In case (b), choose a weakly $\bar\theta$–consistent sequence $\widehat\theta_n$ of G_n–estimators (which exists by Theorem 2.5), thus $\mathbb{P}(B_{\varepsilon,n}) \to 1$, where $B_{\varepsilon,n} = D_n \cap \{d(\widehat\theta_n, \bar\theta) \leq \varepsilon\}$. Moreover $d(\widehat\theta^\rho_n, \rho) \leq d(\widehat\theta_n, \rho)$ by (2.6), hence $\widehat\theta^\rho_n \in M_\varepsilon$ on the set $B_{\varepsilon,n}$, where M_ε is the compact set $\{\theta \in \Theta :\, d(\theta,\rho) \leq d(\bar\theta,\rho) + \varepsilon\}$. Then (6.5) and (6.6) with $M = M_\varepsilon$ and $\widehat\theta^\rho_n$ imply $\mathbb{P}(d(\widehat\theta^\rho_n, \bar\theta) > \varepsilon) \to 0$, hence the weak $\bar\theta$–consistency of $\widehat\theta^\rho_n$. Finally, in both cases, the eventual uniqueness follows from Theorem 2.5. □

Proof of Theorem 2.9. Since each variable $a_n|\widehat\theta_n - \bar\theta|$ is finite–valued (recall the convention that $|\widehat\theta_n - \bar\theta| = 1$ when $\widehat\theta_n = \delta$), it is well known that (2.8) is in fact equivalent to the apparently weaker requirement that

$$\lim_{c\to\infty} \;\limsup_n \mathbb{P}\left(|a_n(\hat\theta_n - \bar\theta)| \leq c\right) = 1. \tag{6.7}$$

We take for $\widehat\theta_n$ the estimator constructed in Theorem 2.5, and in the following we use the notation of the proof of this theorem. Condition 2.8 yields that for any $\varepsilon > 0$ there is a $K > 0$ such that $\mathbb{P}(|G_n(\bar\theta)| > K\Lambda/a_n) \leq \varepsilon$ for all n. Then (6.3) and the fact that $Y(\eta) \to 0$ as $\eta \to 0$ and that $Z_n \to 0$ as $n \to \infty$, in probability, yield that

$$\liminf_n \mathbb{P}(C_{n,K/a_n}) \geq 1 - \varepsilon. \tag{6.8}$$

Moreover, with the sequence ε_n constructed in part 3 of the proof of Theorem 2.5, we know that on C_{n,ε_n} we have $|\widehat\theta_n - \bar\theta| \leq \varepsilon_n$. Hence by (6.1) we necessarily have $|\widehat\theta_n - \bar\theta| \leq K/a_n$ as well on the intersection $C_{n,\varepsilon_n} \cap C_{n,K/a_n}$. Then, combining (6.2) and (6.8), we deduce that

$$\liminf_n \mathbb{P}(|\widehat\theta_n - \bar\theta| \leq K/a_n) \geq 1 - \varepsilon$$

and, since $\varepsilon > 0$ is arbitrary, (6.7) readily follows.

Finally, the last claim follows from the eventual uniqueness proved in Theorem 2.5. □

Proof of Theorem 2.11. By shrinking M if necessary, we may suppose that M is convex. Define $C_n = \{\widehat\theta_n \in M\}$. By the weak $\bar\theta$-consistency we have $\mathbb{P}(C_n) \to 1$, and the mean value theorem yields that on C_n

$$G_n(\widehat\theta_n) - G_n(\bar\theta) = \partial_\theta \widetilde G_n(\widehat\theta_n - \bar\theta).$$

Here $\partial_\theta \widetilde{G}_n$ is the $p\times p$–matrix whose jkth entry is $\partial_\theta G_n(\theta_n^{(j)})_{jk}$, where each $\theta_n^{(j)}$ is a (random) convex combination of $\widehat{\theta}_n$ and $\bar{\theta}$. Observe that

$$\|A_n\,\partial_\theta \widetilde{G}_n\,A_n^{-1} - W(\bar{\theta})\| \le p\left(\sup_{\theta\in M}\|\,A_n\,\partial_\theta G_n(\theta)\,A_n^{-1} - W(\theta)\| + \sup_{\theta:\,|\theta-\bar{\theta}|\le|\widehat{\theta}_n-\bar{\theta}|}\|W(\theta)-W(\bar{\theta})\|\right)$$

on C_n. Hence $A_n\,\partial_\theta\widetilde{G}_nA_n^{-1} \stackrel{\mathbb{P}}{\longrightarrow} W(\bar{\theta})$ because W is a.s. continuous and $\widehat{\theta}_n$ is weakly $\bar{\theta}$-consistent. Therefore, on the set $C_n\cap\{A_n\,\partial_\theta\widetilde{G}_nA_n^{-1}$ is invertible$\}$ (the probability of which goes to 1) we have $G_n(\widehat{\theta}_n)=0$ and thus

$$\begin{aligned} A_n(\widehat{\theta}_n-\bar{\theta}) &= -(A_n\,\partial_\theta\widetilde{G}_nA_n^{-1})^{-1}\,A_nG_n(\bar{\theta}) \\ &= -W(\bar{\theta})^{-1}\,A_nG_n(\bar{\theta}) + U_nA_nG_n(\bar{\theta}), \end{aligned}$$

where $U_n \stackrel{\mathbb{P}}{\longrightarrow} 0$. In view of (2.9), this yields (2.12), and since the convergence is stable and $A_n\,\partial_\theta G_n(\widehat{\theta}_n)\,A_n^{-1} \stackrel{\mathbb{P}}{\longrightarrow} W(\bar{\theta})$, (2.13) also follows. □

Proof of Theorem 3.2. Define $G(\theta)=Q_r(g(\theta))$. Our hypotheses on g, the dominated convergence theorem (implying in particular $\partial_\theta G(\theta)) = Q_r(\partial_\theta g)$ when $\theta\in N$) and (3.1) clearly yield all requirements of Condition 2.2, except (2.2).

For proving (2.2), we first deduce from (3.1) and $\partial_\theta G(\theta)=Q_r(\partial_\theta g)$ that, for any $\theta\in N$,

$$\partial_\theta G_n(\theta) \stackrel{\mathbb{P}}{\longrightarrow} \partial_\theta G(\theta). \tag{6.9}$$

Next, we define for $\eta>0$ a function k_η on D^r and a real number α_η by

$$\begin{aligned} k_\eta(x_1,\dots,x_r) &= \sup_{\theta,\theta'\in M:|\theta'-\theta|\le\eta}\|\partial_\theta g(x_1,\dots,x_r;\theta') - \partial_\theta g(x_1,\dots,x_r;\theta)\| \\ \alpha_\eta &= \sup_{\theta,\theta'\in M:|\theta'-\theta|\le\eta}\|\partial_\theta G(\theta') - \partial_\theta G(\theta)\|. \end{aligned}$$

By (2) of Condition 3.1 the functions $\theta\mapsto\partial_\theta g(s_1,\dots,x_r;\theta)$ are uniformly continuous on the compact set M, hence (3) of this Condition and the dominated convergence theorem yield $Q_r(k_\eta)\to 0$ as $\eta\to 0$. The function $\partial_\theta G(\theta)$ is continuous on the compact set M, so $\alpha_\eta\to 0$.

By the finite covering property of the compact set M, for any $\eta>0$, we have a partition of $l(\eta)$ nonempty subsets M_j of M with diameters less than η. For each $j=1,\dots,l(\eta)$ we choose a point $\theta_j\in M_j$, and we set

$$B_\eta^n = \sum_{j=1}^{l(\eta)}\|\partial_\theta G_n(\theta_j)-\partial_\theta G(\theta_j)\|, \qquad Z_\eta^n = \frac{1}{n}\sum_{i=r}^{n}k_\eta(X_{i-r+1},\dots,X_i).$$

Recalling that $\partial_\theta G_n(\theta)$ equals $\frac{1}{n}\sum_{i=r}^n\partial_\theta g(X_{i-r+1},\dots,X_i;\theta)$, we see that

$$\theta\in M_j \;\Rightarrow\; \|\partial_\theta G_n(\theta)-\partial_\theta G(\theta)\| \le \|\partial_\theta G_n(\theta_j)-\partial_\theta G(\theta_j)\| + Z_\eta^n + \alpha_\eta,$$

hence

$$\sup_{\theta\in M}\|\partial_\theta G_n(\theta)-\partial_\theta G(\theta)\| \le B_\eta^n + Z_\eta^n + \alpha_\eta, \tag{6.10}$$

By (3.1) and (6.9) we have $Z_\eta^n \stackrel{\mathbb{P}}{\longrightarrow} Q_r(k_\eta)$ and $B_\eta^n\stackrel{\mathbb{P}}{\longrightarrow}0$, both as $n\to\infty$ and for any fixed η. Thus, for any $\varepsilon>0$ we can choose first $\eta>0$ and secondly n_0 large enough to have

$$\alpha_\eta\le\varepsilon, \qquad Q_r(k_\eta)\le\varepsilon, \qquad n\ge n_0 \;\Rightarrow\; \mathbb{P}(B_\eta^n>\varepsilon)+\mathbb{P}(Z_\eta^n>Q_r(k_\eta)+\varepsilon)\le\varepsilon.$$

From this and (6.10), we readily deduce the following, which gives us (2.2):

$$\mathbb{P}(\sup_{\theta\in M}\|\partial_\theta G_n(\theta)-\partial_\theta G(\theta)\|>4\varepsilon)\ \le\ \varepsilon.$$

The second part of the theorem follows readily from Theorem 2.7.

□

Proof of Theorem 4.2. Let $\widetilde{Q}$ denote the law of $(Y_0^{\theta_0},\dots,Y_{m\Delta}^{\theta_0})$ on D^{m+1}. With

$$g(x;\theta)=\sum_{j=1}^m h(y_{(j-1)},x_j;\theta)\quad\text{for}\ \ x=(y_0,\dots,y_m)\in D^{m+1}$$

we have $G_n(\theta)=\frac{1}{n}\sum_{i=1}^n g(X_i;\theta)$. Our hypotheses readily imply that g satisfies Condition 3.1 with $N=\Theta$, $\bar\theta=\theta_0$, $r=1$ and $Q_1=\widetilde{Q}$. The X_i's are i.i.d. and hence satisfy (3.1) with almost sure convergence. Thus the existence of strongly consistent, eventually unique G_n–estimators and the last part of the theorem follow from Theorem 3.2 and the comments which follow it. Note that (2.2) for a compact neighbourhood M of θ_0 also follows, with $G(\theta)=\widetilde{Q}(g(\theta))=mQ(h(\theta))$.

The central limit theorem for i.i.d. variables yields (2.9) with $\bar\theta=\theta_0$ and $A_n=\sqrt{n}\,I_p$ (I_p is the identity matrix), where Z is independent of $\mathcal{F}$ and centered Gaussian with covariance matrix $V(\theta_0)=\widetilde{Q}(g(X_1;\theta_0)g(X_1;\theta_0)^*)$, which by (4.1) is equal to $mQ(h(\theta_0)h(\theta_0)^*)$. Moreover (2.2) and $A_n=\sqrt{n}\,I_p$ imply (2.10) with $W(\theta)=\partial_\theta G(\theta)=mQ(\partial_\theta h(\theta))$. Now (4.2) follows from Theorem 2.11, and the last statements follow from Theorem 2.7. □

Proof of Theorem 5.2. 1) Below, $M\subseteq B$ is a fixed compact subset, and C is a generic constant. In view of Assumption 5.1 and of (5.2), the following estimates are classical (and easy to derive: use Itô's formula applied to X^2W and to $c(X;\beta_0)$ for the third one).

$$\begin{aligned}
&|\,\mathbb{E}((\Delta_i^nX)^2\mid\mathcal{F}_{(i-1)\Delta_n})-\Delta_n c(X_{(i-1)\Delta_n};\beta_0)|\le C\Delta_n^2(1+|X_{(i-1)\Delta_n}|^C)\\
&\mathbb{E}(|\Delta_i^nX|^{2m}\mid\mathcal{F}_{(i-1)\Delta_n})\le C_m\Delta_n^m(1+|X_{(i-1)\Delta_n}|^{C_m})\\
&|\,\mathbb{E}((\Delta_i^nX)^2\,\Delta_i^nW\mid\mathcal{F}_{(i-1)\Delta_n})|\le C\Delta_n^2(1+|X_{(i-1)\Delta_n}|^C).
\end{aligned}$$

Then, setting

$$\begin{aligned}
f(x,\beta)&=\frac{\partial_\beta c(x;\beta)}{c(x;\beta)^2}(c(x;\beta_0)-c(x;\beta))\\
\zeta(\beta)_i^n&=\frac{\partial_\beta c(X_{(i-1)\Delta_n};\beta)}{c(X_{(i-1)\Delta_n};\beta)^2}((\Delta_i^nX)^2-\Delta_n c(X_{(i-1)\Delta_n};\beta)),
\end{aligned}$$

($\zeta(\beta)_i^n$ is the ith summand in (5.3)), we deduce for all $\beta\in M$ and $m=0,1$:

$$\Big|\,\mathbb{E}(\partial_\beta^m\zeta(\beta)_i^n\mid\mathcal{F}_{(i-1)\Delta_n})-\Delta_n\partial_\beta^m f(X_{(i-1)\Delta_n};\beta)\Big|\le C\Delta_n^2(1+|X_{(i-1)\Delta_n}|^C)\tag{6.11}$$

$$\mathbb{E}(|\partial_\beta^m\zeta(\beta)_i^n|^2\mid\mathcal{F}_{(i-1)\Delta_n})\le C\Delta_n^2(1+|X_{(i-1)\Delta_n}|^C)\tag{6.12}$$

$$\mathbb{E}(\sup\nolimits_{\beta\in M}|\partial_\beta^2\zeta(\beta)_i^n|\mid\mathcal{F}_{(i-1)\Delta_n})\le C\Delta_n(1+|X_{(i-1)\Delta_n}|^C)\tag{6.13}$$

$$\Big|\,\mathbb{E}((\zeta(\beta_0)_i^n)^2\mid\mathcal{F}_{(i-1)\Delta_n})-2\Delta_n^2\frac{(\partial_\beta c(X_{(i-1)\Delta_n};\beta_0))^2}{c(X_{(i-1)\Delta_n};\beta_0)^2}\Big|\le C\Delta_n^3(1+|X_{(i-1)\Delta_n}|^C)\tag{6.14}$$

$$\mathbb{E}(|\zeta(\beta_0)_i^n|^4\mid\mathcal{F}_{(i-1)\Delta_n})\le C\Delta_n^4(1+|X_{(i-1)\Delta_n}|^C)\tag{6.15}$$

$$\Big|\,\mathbb{E}(\zeta(\beta_0)_i^n\Delta_i^nW\mid\mathcal{F}_{(i-1)\Delta_n})\Big|\le C\Delta_n^2(1+|X_{(i-1)\Delta_n}|^C)\tag{6.16}$$

2) Define G and W by (5.4) and (5.5). Then $G(\beta_0) = 0$, the C^1 property of G_n and G are obvious, $W(\beta) = \partial_\beta G(\beta)$, and the identifiability condition implies $W(\beta_0) < 0$ a.s. We have $\Delta_n = T/n$, hence (6.11) and (6.12) for $m = 0, 1$ plus Riemann integration yield $G_n(\beta_0) \xrightarrow{\mathbb{P}} 0$ and $\partial_\beta G_n(\beta) \xrightarrow{\mathbb{P}} W(\beta)$. Inequality (6.13) implies that $\beta \mapsto \partial_\beta G_n(\beta))$ is Lipschitz on M with a (random) Lipschitz coefficient that is integrable. This together with the pointwise in β convergence in probability implies by standard tightness arguments that (2.2) and (2.10) with $A_n = \sqrt{n}$ hold.

By Theorem IX.7.28 of Jacod and Shiryaev (2003) plus again Riemann integration, (2.9) is a straightforward consequence of (6.11) with $m = 0$ and $\beta = \beta_0$, plus (6.14)–(6.16). Hence, we have proved that Conditions 2.2 and 2.10 hold, and the last claim follows from Theorem 2.7. □

Proof of Theorem 5.5. Let us denote by $\zeta'^j(\theta)_i^n$ the ith summands in the definition (5.7) of $G_n^j(\theta)$, $j = 1, 2$. We also define for $\theta = (\alpha, \beta)$

$$\widetilde{f}(x, \theta) = \frac{\partial_\alpha a(x; \alpha)}{c(x, \beta)} \, (a(x; \alpha_0) - a(x; \alpha))$$

and observe that

$$|\, \mathbb{E}(\Delta_i^n X \mid \mathcal{F}_{(i-1)\Delta_n}) - \Delta_n a(X_{(i-1)\Delta_n}; \alpha)| \leq C\Delta_n^2(1 + |X_{(i-1)\Delta_n}|^C).$$

Then for all $\theta \in M$ (with M a compact subset of Θ) and $m = 0, 1$ we have

$$\Big|\, \mathbb{E}(\partial_\theta^m \zeta'^1(\theta)_i^n \mid \mathcal{F}_{(i-1)\Delta_n}) - \frac{\Delta_n}{T_n} \, \partial_\theta^m \widetilde{f}(X_{(i-1)\Delta_n}; \theta)\Big| \leq C\frac{\Delta_n^2}{T_n} \, (1 + |X_{(i-1)\Delta_n}|^C)$$

$$\mathbb{E}(|\partial_\theta^m \zeta'^1(\theta)_i^n|^2 \mid \mathcal{F}_{(i-1)\Delta_n}) \leq C\frac{\Delta_n}{T_n^2} \, (1 + |X_{(i-1)\Delta_n}|^C)$$

$$\mathbb{E}(\, \sup_{\theta \in M} \, |\partial_\theta^2 \zeta'^1(\theta)_i^n| \mid \mathcal{F}_{(i-1)\Delta_n}) \leq C\frac{\Delta_n}{T_n} \, (1 + |X_{(i-1)\Delta_n}|^C)$$

$$\Big|\, \mathbb{E}((\zeta'^2(\theta_0)_i^n)^2 \mid \mathcal{F}_{(i-1)\Delta_n}) - \frac{\Delta_n}{T_n^2} \, \frac{(\partial_\alpha a(X_{(i-1)\Delta_n}; \alpha))^2}{c(X_{(i-1)\Delta_n}; \beta)}\Big| \leq C\frac{\Delta_n^2}{T_n^2} \, (1 + |X_{(i-1)\Delta_n}|^C)$$

$$\mathbb{E}(|\zeta'^1(\theta_0)_i^n|^4 \mid \mathcal{F}_{(i-1)\Delta_n}) \leq C\frac{\Delta_n^2}{T_n^4} \, (1 + |X_{(i-1)\Delta_n}|^C)$$

$$|\, \mathbb{E}((\zeta'^1(\theta_0)_i^n \zeta'^2(\theta_0)_i^n \mid \mathcal{F}_{(i-1)\Delta_n})| \leq C\frac{\Delta_n^2}{T_n^2} \, (1 + |X_{(i-1)\Delta_n}|^C).$$

Since $\zeta'^2(\theta)_i^n = \zeta(\beta)_i^n / T_n$, the variables $\zeta'^1(\theta)_i^n$ satisfy (6.11)–(6.16), after normalization by an appropriate power of T_n.

At this stage, the proof follows the scheme of the previous proof, except that we use the law of large numbers (5.6) instead of the convergence of Riemann sums (we leave the – tedious – details to the reader). In particular, (2.9) is a consequence of classical convergence results, for which we do not need the analogue of (6.16) because in the ergodic case the limit Z in (2.9) is automatically independent of $\mathcal{F}$. Hence, Conditions 2.2 and 2.10 hold, and the last claim follows from Theorem 2.7. □

References

Aitchison, J. and Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.*, **29**, 813–828.

Barndorff-Nielsen, O. E. and Sørensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes. *International Statistical Review*, **62**, 133–165.

Bibbona, E. and Ditlevsen, S. (2012). Estimating in discretely observed diffusions killed at a threshold. *Scand. J. Statist.*, **40**, 274–293.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34**, 305–34.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data, 2nd Edition*. Oxford University Press, Oxford.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *J. Roy. Statist. Soc. Ser. B*, **22**, 139–153.

Fisher, R. A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.*, **98**, 39 – 54.

Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales Inst. H. Poincaré (B)*, **29**, 119–151.

Gobet, E. (2001). Local asymptotic mixed normality property for elliptic diffusion: a malliavin calculus approach. *Bernoulli*, **7**, 899–912.

Gobet, E. (2002). Lan property for ergodic diffusions with discrete observations. *Annales Inst. H. Poincaré (B)*, **38**, 711–737.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208–1212.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika*, **72**, 419–428.

Godambe, V. P. and Heyde, C. C. (1987). Quasi likelihood and optimal estimation. *International Statistical Review*, **55**, 231–244.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press, New York.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics*, **30**, 203–238.

Hansen, L. P. (2001). Method of moments. In *International Encyclopedia of the Social and Behavior Sciences*. New York: Elsevier.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer-Verlag, New York.

Jacod, J. and Shiryaev, A. (2003). *Limit Theorems for Stochastic Processes, 2d ed.* Springer Verlag, Berlin.

Jakobsen, N. M. and Sørensen, M. (2017). Efficient estimation for diffusions sampled at high frequency over a fixed time interval. *Bernoulli*, **23**, 1874–1910.

Kimball, B. F. (1946). Sufficient statistical estimation functions for the parameters of the distribution of maximum values. *Ann. Math. Statist.*, **17**, 299–309.

Kuersteiner, G. (2002). Efficient instrumental variables estimation for autoregressive models with conditional heteroskedasticity. *Econometric Theory*, **18**, 547–583.

Loomis, L. H. and Sternberg, S. (1968). *Advanced Calculus*. Addison-Wesley, Reading, MA.

Pedersen, A. (2000). Estimating the nitrous oxide emission rate from the soil surface by means of a diffusion model. *Scand. J. of Statistics*, **27**, 385–403.

Sørensen, M. (1999). On asymptotics of estimating functions. *Brazilian Journal of Probability and Statistics*, **13**, 111–136.

Sørensen, M. (2017). Efficient estimation for ergodic diffusions sampled at high frequency. Preprint, University of Copenhagen, Denmark.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.*, **8**, 1375–1381.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *J. Multivariate Anal.*, **65**, 245–260.